图书情报工作
第63卷 第5期 2019年3月

基于深度学习的MOOC论坛探索型对话识别方法研究*

董庆兴[1,2] 李华阳[3] 曹高辉[1] 夏立新[1]
[1] 华中师范大学信息管理学院 武汉 430079 [2] 武汉大学信息资源研究中心 武汉 430079
[3] 腾讯 AI Lab 深圳 518057

摘要：[目的/意义]大规模在线开放课程论坛具有丰富的用户评论数据。从大量未区分的评论数据中，自动识别出知识密度较高的探索型对话并挖掘其潜在价值，对于改善教师教学质量以及提高学生知识水平具有重要影响。[方法/过程]首先利用 GloVe 方法训练词向量，加强对文本语义的理解，然后利用卷积神经网络自动学习文本特征，提出一种基于深度学习的探索型对话自动识别模型，并在学堂在线平台《心理学概论》课程论坛标注数据集上进行实证与对比研究。[结果/结论]实验结果显示，利用 GloVe 方法预训练词向量以及在训练过程中不断对词向量进行学习修正能够提高模型效果。该模型识别探索型对话的 F1 值为 0.94，相较于传统的朴素贝叶斯方法(0.88)、逻辑斯谛回归方法(0.89)、决策树方法(0.88)以及随机森林方法(0.88)取得较大提升，具有较高的实用性和较低的学习成本。

关键词： MOOC论坛 探索型对话 GloVe 卷积神经网络
分类号： G251
DOI： 10.13266/j.issn.0252-3116.2019.05.011

1 引言

在当前教育大数据背景下，大规模在线开放课程[1]（Massive open online course，MOOC）论坛中知识密度较高的探索型对话[2-3]是改善 MOOC 教育质量的重要着力点。随着大数据时代的来临，教育大数据正在成为一种变革教育管理的战略资产和科学力量[4]。利用人工智能和数据挖掘等技术对教育数据进行分析，可以帮助教育工作者确定教育资源分配政策、识别有辍学倾向的学生、加强教学过程干预等[5]。近年来，MOOC 在教学过程中产生了大量用户参与、学习与反馈的数据，是一类重要的数据来源[6-7]。MOOC 以互联网为基础，学生参与学习可以不受时空上的限制，但这也使得参与者无法像传统课堂一样进行面对面的交流，因此课堂参与者的主要对话形式变为 MOOC 论坛这种虚拟社区[8]中的发帖和回复。然而，外在形式的不同并没有影响课堂对话以信息和知识交流为目的的本质[9]。作为课堂中具有大量知识产出的高效对话方式，探索型对话是一种参与者利用语言进行共同推理，以合理平等的方式分享知识、挑战观点、评估证据、考察备选方案并最终达成共识的对话形式。在 MOOC 论坛中，能够帮助用户将大脑中的隐性知识编码为显性知识的探索型对话对于构建良好的虚拟社区生态环境、推动知识共享[10]等具有重要作用，因此以探索型对话为依托来改善 MOOC 的教育质量具有现实意义。

当前在参与者进行信息交流的过程中，MOOC 论坛积累了大量内容相对集中的文本对话，但是其中大部分是以结交朋友、询问考试时间、询问作业能否延期提交等为主要内容的课程知识密度含量较低的非探索型对话，对于提高教育质量影响较小。主流 MOOC 平台普遍未对探索型对话和非探索型对话加以区分，从而导致数量相对较少且发布时间分散的探索型对话淹没在大量非探索型对话当中，难以有效发挥其在 MOOC 课程中帮助改善教育教学的作用。虽然绝大多数论坛都具备检索功能，但是教师和学生利用检索的

* 本文系国家自然科学基金项目“面向群智感知大数据的群体评价模型与方法研究”(项目编号:71871102)和华中师范大学中央高校基本科研业务费项目“在线医疗服务环境下用户信息感知及其线下行为变化机制研究”(项目编号:CCNU17TS0009)研究成果之一。

作者简介：董庆兴(ORCID:0000-0003-3512-9333)，系主任，副教授，博士；李华阳(ORCID:0000-0003-3539-8648)，助理研究员，本科；曹高辉(ORCID:0000-0002-2760-4889)，副院长，副教授，博士，通讯作者，E-mail:ghcao@mail.ccnu.edu.cn；夏立新(ORCID:0000-0002-4162-2282)，副校长，教授，博士。

收稿日期:2018-06-27 修回日期:2018-08-12 本文起止页码:92-99 本文责任编辑:杜杏叶

方法只能查找到符合特定关键词的对话,无法像识别探索型对话一样在论坛中对所有探索型对话进行聚合,不利于学生的参与和教师管理。而识别探索型对话,有利于论坛参与者发现探索型对话并积极响应。在这一过程中,教师可以对学习者加以引导,激发他们深入思考,从而在交流过程中获得更多的知识。

识别探索型对话的方法主要包括人工标注和自动识别两种,人工标注的方法虽然准确率较高但是随着用户量的增加和论坛规模的扩大,这种费时费力的管理方法会加重论坛管理者的负担,已经不再适用于大规模论坛。许多传统的基于文本分类的探索型对话自动识别方法由于理解语义以及学习特征的能力不足在最终的识别准确率上表现较差,因此本文针对性的利用 GloVe 词向量[11]加强语义理解能力,以及利用模型学习能力较强的卷积神经网络[12](Convolutional neural networks,CNN)来共同提高探索型对话识别的准确性,以便将来在真实场景中更好地对探索型对话进行深入研究和应用。

2 相关研究回顾

文本分类作为信息科学领域的一个经典课题,学者们已经开展了一系列面向不同具体任务的方法研究[13],尤其在词语表示以及特征工程上进行了大量探索。在词语编码方面,大多数的文本分类研究利用 $R^{|V|\times 1}$ 维的向量空间模型[14](Vector space model, VSM)对词语进行表示,其中 $|V|$ 表示语料库中词典的大小,向量空间中的每一维度对应一个单词。根据单词对应维度的特征值的计算方式的不同,常用的编码方式主要有,One-hot、tf 以及 tf-idf。One-hot 编码的表示向量中,每一维度的特征值均属于集合{0,1},如 M. Shahami 等[14]利用 One-hot 方法对邮件中的单词进行编码时,1 表示该单词在邮件中出现,0 表示未出现。虽然 One-hot 方法在一些场景下表现出了良好的效果,但是却忽略了文章词语之间的差异信息,因此在文本分类领域经常利用 tf 或者 tf-idf 算法计算单词对应维度的特征值[15-16],在一定程度上表征单词在文本中的不同重要程度。虽然 tf-idf 算法在信息检索和自然语言处理领域得到了广泛应用,仍然有研究针对 tf-idf 存在的类区分能力不足等问题进行后续改进[17]。

在传统文本分类研究领域中,为了提高特定场景下的文本分类效果,通常需要结合大量特征工程的方法。为了提高先秦诸子典籍的自动分类效果,王东波等[18]分别对比了 tf-idf、信息增益、互信息和卡方分布 4 种方法抽取的特征对分类器性能的影响,结果发现在特征维数相同时,利用 tf-idf 方法抽取特征训练的分类器明显优于其他方法。针对信息增益算法试用范围较广,但是精度较低的问题,夏火松等[19]提出一种基于领域词典结合评论长度特征提取方法,该方法改进了普通领域词典建立耗时长的问题,并且结合评论长度解决了词典难以跨领域的问题。为了更好地筛选特征、降低特征维度、控制特征稀疏性,杜亚楠等[20]引入了一种基于 G^2 检验的特征筛选方法,使用该方法能够获得强区分能力的特征。通过抽取字符特征、词汇特征、句法特征和文本布局特征等,祁瑞华等[21]构建的多层面的文体风格特征模型能够较好地适应短文本,具有较高的鲁棒性,在博客作者识别任务上取得了出色的效果。考虑到语言网络也是一种典型的复杂网络,李晓军等[22]运用复杂网络中的研究方法构建文本网络,引入词语之间的最短路径等网络特征来提高作者身份识别的准确率。

虽然上述研究在具体任务中都取得了不错的效果,但是还存在一些值得改进之处:首先,在 $R^{|V|\times 1}$ 维的向量空间模型中,所有单词的表示向量都相互正交,词与词之间的关系难以衡量,对语义的理解较差[23]。此外,向量维度会随着语料库中单词的数量增加而增加,当语料过多时会导致维度爆炸[24]。不同于之前的向量空间模型的编码方式,词向量的主要思想则是利用语言学信息在低维子空间中学习出一种对词语语义有较好理解的表示向量,且表征向量的维度不会随着词典大小的增加而变化。因此,利用词向量有助于解决传统编码方法语义表达能力不足以及维度灾难的问题。此外,许多文本分类模型的学习能力较弱,往往需要人为构造复杂的特征才能达到较好的效果。而深度学习方法[25]能够自动学习数据中的特征,尤其是已经在很多自然语言处理(Natural Language Processing, NLP)任务中取得了很好效果[26-27]的卷积神经网络能够利用卷积核学习词组片段更抽象的表示,从而捕捉文本中更深层次的特征,使得模型在不需要复杂的特征工程的情况下取得出色效果。

3 探索型对话自动识别模型

本文主要从语义编码以及特征学习上进行改进,首先利用 GloVe 模型预训练词向量,在连续的低维空间对词语进行编码加强对词语语义的理解,然后利用能够自动学习文本深层次特征的卷积神经网络自动识别探索型对话,以解决 MOOC 论坛场景下传统文本分

类方法在语义学习和特征学习方面能力不足的问题。

3.1 GloVe 词向量模型

宏观上看主要有两类不同的词向量模型：第一类模型主要依赖矩阵分解[28]，此类模型能够有效利用单词共现来捕捉单词之间的相似度，但是在同义词类比等任务上效果较差；第二类模型主要基于浅层窗口[29-30]，利用 Skip-gram 或 CBOW 等模型[31]学习语料中复杂的语言学模式，但是这类模型并没有利用全局统计信息。而 GloVe 词向量利用加权平均损失模型在单词共现矩阵上进行训练，能够更好地利用全局统计信息，相比于其他词向量方法在词语相似度以及 NER 等自然语言处理的下游任务上表现出更好的效果[11]。因此在 MOOC 场景下的探索型对话识别任务上，本文将采用 GloVe 模型预训练单词的词向量。

在利用 GloVe 模型训练词向量之前，首先要遍历整个语料库并统计出大小为 $|V| \times |V|$ 的共现矩阵 X。其中矩阵 X 的元素 X_{ij} 表示单词 i 和 j 共同出现在一个窗口内的次数，$X_i = \sum_k X_{ik}$ 表示单词 i 出现的总次数。由于单词的语义很大程度上是由单词的上下文所表达的[31]，单词 j 在单词 i 的上下文中出现的概率 $p(j|i) = \frac{X_{ij}}{X_i}$ 越大表示单词 i 和 j 之间的语义联系越紧密。例如“冰”和“水蒸气”同是水的两种状态，而由于“固体”和冰在语义上的共性要远大于“固体”和“水蒸气”的语义共性，所以单词“固体”在“冰”的上下文中出现的概率会远高于在单词“水蒸气”的上下文中出现概率[11]。共现矩阵 X 的统计结果是对单词之间语义关系的一种反应，GloVe 模型的主要目的是为了学习出一种词向量的编码方式，使得利用该方式编码的词向量计算出的概率分布 Q 尽可能的逼近矩阵 X 的分布。GloVe 模型的损失函数 J 采用交叉熵损失函数，如式(1)所示：

$$J = -\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} X_{ij}\log Q_{ij} \qquad 式(1)$$

其中，

$$Q_{ij} = \frac{\exp(u_j^T v_i)}{\sum_{w=1}^{|V|}\exp(u_w^T v_i)} \qquad 式(2)$$

式(2)中，u_j 表示单词 j 作为输出时的词向量，v_i 表示单词 i 作为输入时的词向量。

Q_{ij} 的计算开销是非常高的，因为在 Q_{ij} 的分母部分，每一输入 v_i 都需要和词表中的所有单词计算一次相似度以保证 $\sum_j Q_{ij} = 1$。为了减少计算概率分布 Q 时标准化的昂贵开销，GloVe 改用最小二乘损失函数如式(3)所示：

$$\hat{J} = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} X_i(\hat{P}_{ij} - \hat{Q}_{ij})^2 \qquad 式(3)$$

其中 $\hat{P}_{ij} = X_{ij}$ 并且 $\hat{Q}_{ij} = \exp(u_j^T v_i)$。

此外，对于大型语料库，X_{ij} 过大可能导致模型训练困难，为解决这一问题，GloVe 模型对 $\hat{P}$ 和 $\hat{Q}$ 分别取对数，同时利用 $f(x)$ 在损失函数中对 X_i 这一加权项进行映射：

$$J = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} f(X_i)\log(\hat{P}_{ij} - \log(\hat{Q}_{ij})^2) \qquad 式(4)$$

其中：

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \qquad 式(5)$$

式(5)中的 x_{max} 和 α 属于超参数。

3.2 卷积神经网络模型

卷积神经网络对于文本序列表征向量层次化的建模方式来说能够学习到更深层次特征。如图 1 所示，第一层是文本序列的输入层，在第二层卷积神经网络利用 $w \in R^{k \times d}$ 的卷积核对句子进行卷积能够学习出句子中 k - grams 对应的特征值[32]，其中 k - grams 表示文本序列中所有 k 个连续单词构成的序列片段，每个序列片段称为一个 gram。卷积核学习出的 k - grams 对应的表征向量实际上是对文本序列更细粒度的表示，相比于直接在句子层面对文本序列进行建模，卷积神层网络能够捕获更深层次的语言学信息和有效特征。在最后一层卷积神经网络会对 k - grams 的表征向量做更深层次的聚合，并最终得到文本序列在句子层次的表示。

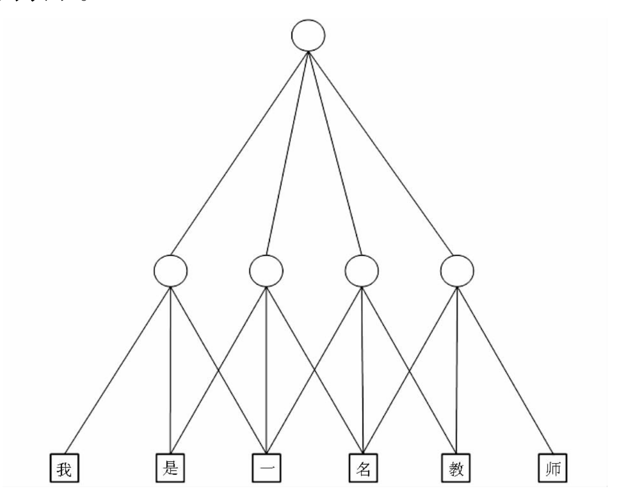


图 1 卷积神经网络特征学习原理图

探索型对话自动识别模型的结构见图 2。经数据预处理的原始文本传入词嵌入层之后得到 CNN 模型的输入矩阵 $M \in R^{n \times d}$。其中 n 表示对话中单词的个数，d 表示词向量的维度，M_i 表示单词 i 对应的词向量。

嵌入层之后的卷积 – 池化层是整个 CNN 模型的核心，为了使模型更好地学习文本特征，本文采用了三个卷积核大小不同的卷积池化层，其详细结构见图 3。

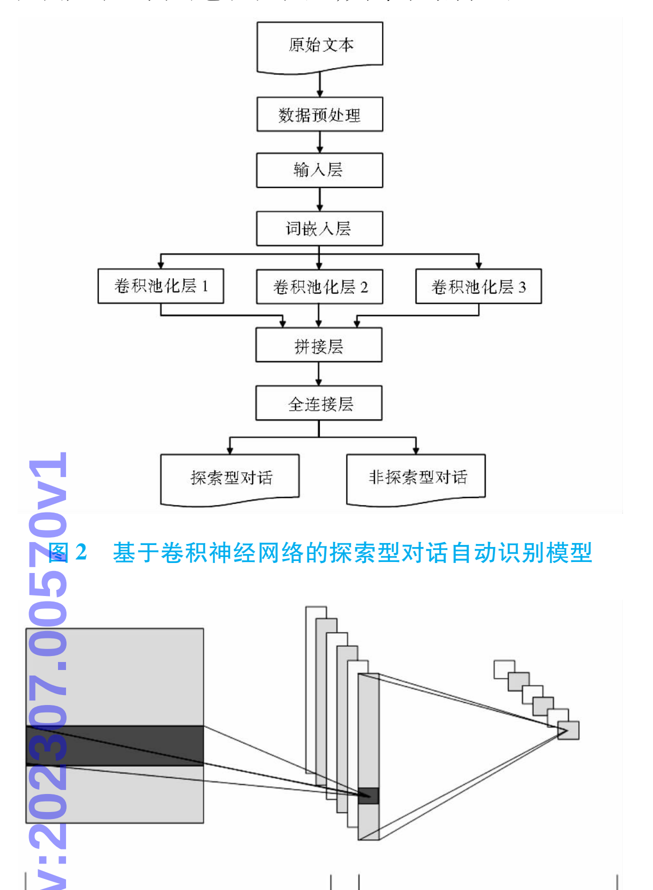


图 2　基于卷积神经网络的探索型对话自动识别模型

图 3　卷积池化层结构

图 3 中，卷积 – 池化层主要由两部分组成：卷积层和池化层。对于 $W \in R^{k \times d}$ 的卷积核，经卷积操作得到新的特征 c_i：

$$c_i = f(w \cdot M_{i:i+k-1} + b) \quad \text{式(6)}$$

其中，b 为偏差项，函数 f 为非线性激活函数，例如 sigmoid 函数。c_i 可以理解为对输入文本中单词 i 到 $i+k-1$ 所组成词组的抽象表示。卷积核对输入矩阵 M 卷积结束之后得到 $c = [c_1, c_2, \cdots, c_{n-k+1}] \in R^{n-k+1}$。

考虑到不同卷积核卷积得到的特征向量 c 的维度是不同的，而后续层的输入维度又必须是固定的，所以我们需要对 c 的维度重新进行调整。最大池化（max pooling）是一种流行的解决方案，将 $c \in R^{n-k+1}$ 输入最大池化层将得到 $\hat{c} = \max(c)$。其中 $\hat{c} \in R$，这一方面固定了特征向量的维度，另一方面又保留最重要的特征。

所有卷积 – 池化层的输出特征的维度都相同，这些特征拼接成的特征向量将输入全连接层。为了防止模型过拟合，在模型训练的时候会对全连接层进行 dropout[33] 处理，然后再使用 softmax 函数对全连接层的输出进行预测。为了提高模型的泛化能力，模型的损失函数在使用交叉熵损失函数的基础上，加入了全连接层权重的 l_2 正则项，然后利用 Adam 方法[34] 对损失函数进行优化。

4　研究设计

学堂在线这一中文 MOOC 平台上开设的《心理学概论课程》由于参与人数众多，并且课程论坛的对话多以汉语文本描述为主，成为本文实验数据获取的首选目标。本实验数据集由 PySpider 网络爬虫框架在该课程 2015 年春和 2016 年春的课程论坛上抓取的相关用户数据（用户 ID 等）和发帖数据（标题、时间、内容、回复等）构成。两期课程的数据概况见表 1，2015 年春和 2016 年春课程报名人数分别为 31 680 人和 23 372 人，报名人数下降非常明显。此外，论坛参与者数量相较于课程报名人数比重较小，两期课程论坛参与人数分别为 1 002 人和 274 人。论坛中的文本对话由用户发帖以及用户的互动回复两部分构成，其中用户发帖数量分别为 1 029 个和 221 个，用户回复数量约为发帖数量的 3 倍，分别为 3 165 个和 788 个。数据中的探索型对话由华中师范大学教育大数据应用技术国家工程实验室师生手工标注完成，数量分别为 344 个和 62 个，分别占该年课程论坛中所有对话的 1/3 左右。可以看出探索型对话在 MOOC 论坛中数量较少，这一特点加大了探索型对话的检索难度。总体上看，对比 2015 年和 2016 年两期课程在各项数据上均有明显下降，由于 2016 年论坛数据过于稀少，本文主要以 2015 年数据集为基础进行后续分析和实验，标注后的数据截屏见图 4。

表 1　两期《心理学概论》课程数据概况

课程时间	报名人数	论坛参与人数	对话数量	回复量	探索型对话
2015 年春	31 680	1 002	1 029	3 165	344
2016 年春	23 372	274	221	788	62

ID	探索型对话	标题	内容
54f669a8f605ab098e001d71	1	新生如何学好心理学?	1、心理学可以作为一种兴趣，但是
54f70e8bf605ab58bd0020fd	1	为什么只做出纳员的概率比较高啊?	你们看啊，根据选项，我们可以这
54f677d1f605abdf61001dd0	1	关于心态，学习的心态。	“”分“”别看了国内与国外的两本心理
54f66c27f605ab03c7001d7e	0	请教上课的时间安排	请教上课的具体时间安排
54f670e9459f08b862001dfa	0	讲义怎么下载啊?	搜索云盘，显示已经不在了。
54f674ee459f086fe4001e17	0	课程讲义没法下载，请更新下载链接	如题

图 4　《心理学概论》论坛数据标注样例

2015 年课程中探索型对话在 24 小时内的发帖分布见图 5。在 1 点到 15 点之间，除了 13 点和 14 点出现明显波峰之外，探索型对话的发布整体比较均匀。

图 6 绘制的是探索型对话和非探索型对话随着课程不断推进的发布数量变化趋势，从中可以看出，从课程开始到课程第 80 天左右，探索型对话和非探索型对话的变化趋势相近，但是探索型对话数量相对较少；在课程第 80 天之后，探索型对话基本消失，但是非探索型对话却出现了一次波峰，并且一直到课程第 140 天左右依旧有对话发布，这些对话大多在询问考试成绩等问题，与课程内容无关。综合上述分析，探索型对话确实存在数量较少，发布时间比较分散等规律，这也从侧面反映了实现探索型对话自动识别的必要性。

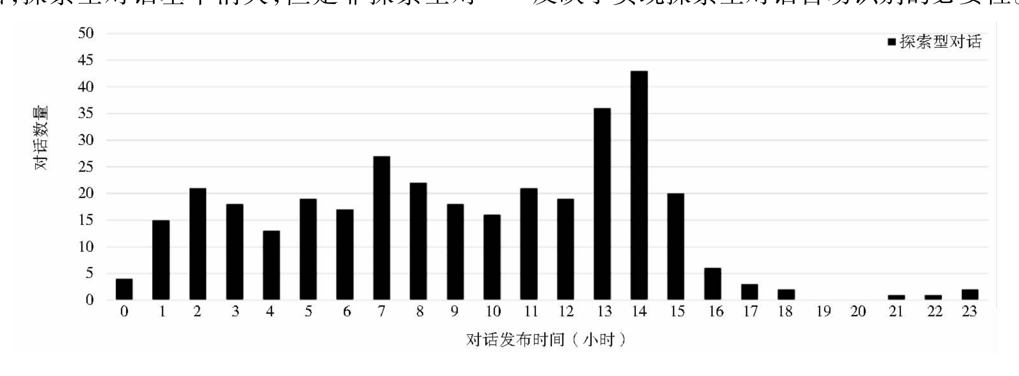


图 5　探索型对话发帖时间的 24 小时分布图

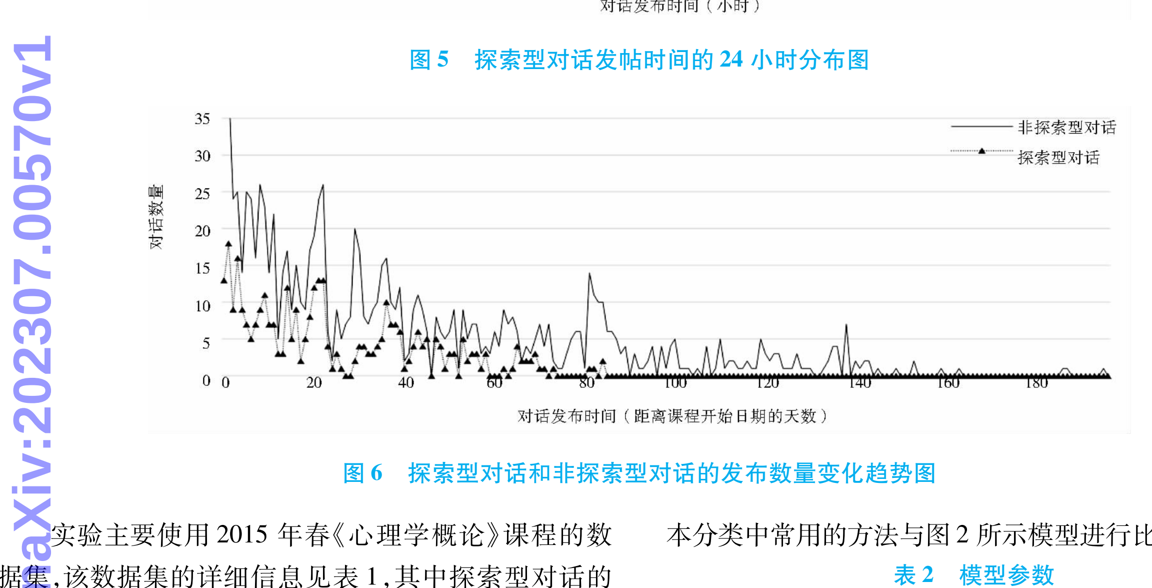


图 6　探索型对话和非探索型对话的发布数量变化趋势图

实验主要使用 2015 年春《心理学概论》课程的数据集，该数据集的详细信息见表 1，其中探索型对话的数量约为非探索型对话数量的，不同种类样本数量的不均衡将导致模型更倾向于将对话识别为非探索型对话。为了平衡两类对话的数量，本研究随机选择出一些探索型对话进行复制，直到两类对话数量相等。平衡数据集之后，需要对文本进行分词，并构建出数据集的单词表，然后将文本中的单词转换成对应的索引。训练时的参数如表 2 所示，通过在整个数据集中随机挑选 90% 得到训练集，剩下的 10% 作为测试集。

实验的总体整体流程设计见图 7，其中文本分类方法主要包括卷积神经网络方法和传统文本分类方法两大类。首先，本文将探究 GloVe 方法预训练的词向量以及随机初始化的词向量对卷积神经网络最终分类效果的影响。其次，为了对比卷积神经网络与传统文本分类方法在探索型对话自动识别任务上的效果，本文选取了朴素贝叶斯模型（Naive Bayes，NB）、逻辑斯谛回归模型（Logistic Regression，LR）、决策树模型（Decision Tree，DT）以及随机森林模型（Random Forest，RF）这些传统文本分类中常用的方法与图 2 所示模型进行比较。

表 2　模型参数

参数	说明
共现矩阵窗宽	GloVe 模型统计矩阵 X 时的窗宽设置为 1
x_{max}	100
α	0.75
卷积核宽度	CNN 模型卷积核的宽度（h）设置为 3，4，5
词向量维度	词向量维度设置为 128
Dropout	Dropout 值设为 0.5
正则项系数	l_2 正则项的系数设为 0.1
优化方法	Adam
Batch-size	Mini-batch 的大小设为 64

对于分类结果，本文采用文本分类领域普遍使用的准确率（accuracy）、精确率（precision）、召回率（recall）以及 F1 值[35]。具体的计算公式如下：

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad 式(7)$$

$$precision = \frac{TP}{TP + FP} \quad 式(8)$$

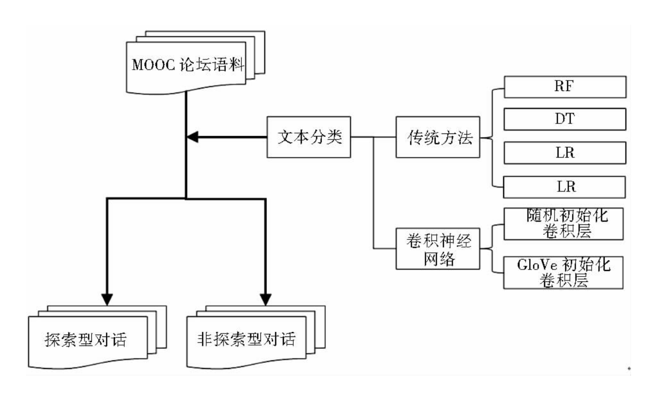


图 7　MOOC 论坛探索型对话自动识别流程图

$$recall = \frac{TP}{TP + FN} \quad \text{式（9）}$$

$$F1 = \frac{2\ precision \times recall}{precision + recall} \quad \text{式(10)}$$

其中,*TP*、*FP*、*TN*、*FN* 分别表示探索型对话被识别为探索型对话的个数、非探索型对话被识别为探索型对话的个数、非探索型对话被识别为非探索型对话的个数以及探索型对话被识别为非探索型对话的个数。

5　结果分析

5.1　词嵌入层对分类结果影响分析

词嵌入层包含了模型输入所需的全部特征,是整个分类器的重要一环,探究词嵌入层不同的初始化和训练方法对最终的探索型对话识别结果的影响具有重要意义。对于表 3 中的初始化方法,预训练的词向量是指前文提到的利用 GloVe 方法预训练得到的结果,均匀分布随机初始化是指利用在［－1，1］区间内的均匀分布对词嵌入层进行随机初始化;对于表 3 中第二列,在训练过程中是否继续调整词嵌入层是指在分类器训练时是否计算词嵌入层参数的梯度并对其进行优化。

在保证卷积神经网络模型参数固定的情况下,仅改变词嵌入层的设置,得到了如表 3 所示的实验结果。实验结果表明 GloVe 模型预训练的词向量是有效的并且在训练过程中继续调整词嵌入层有助于提升模型的分类效果。首先,从表 3 的第一行可以看出,仅使用预训练的词向量模型已经取得了非常好的效果。虽然对比表中的后两行,在不断调整词嵌入层的前提下利用预训练的词向量初始化相较于随机初始化的模型在准确率和精确率上只有微小提升,但这并不意味着模型中独立存在的词向量预训练过程是没有意义的。因为在大部分研究中,词向量往往只是具体任务的一个副产物,目前许多针对如何训练词向量的研究实际上是在寻找能够高效、快速地学习出较好地表达了单词语义的词向量的方法[31]。而在本文所提出的模型中,能更方便地学习出效果较好的词向量,将词向量学习的过程与最终分类器的训练过程分离,可以减少模型的训练负担,这在大规模任务上具有重要意义。其次,对比表 3 中的前两行可以发现,在训练的过程中继续调整词嵌入层使得模型的准确率、召回率以及 F1 值均得到了提高,虽然精确率有所下降,但是综合来看继续调整词嵌入层确实有助于提升模型的整体效果。

表 3　词嵌入层不同初始化和训练方法下的模型分类效果

初始化方法	训练过程继续调整	准确率	精确率	召回率	F1 值
预训练的词向量	否	0.94	0.93	0.96	0.94
预训练的词向量	是	0.95	0.92	0.99	0.95
均匀分布随机初始化	是	0.94	0.91	0.99	0.95

5.2　与传统文本分类模型的对比分析

为了对比本文提出的模型与传统文本分类模型在 MOOC 探索型对话自动识别任务上的效果,如图 6 所示本文选取了在文本分类领域比较经典的几种模型进行了比较。其中传统文本分类方法将采用情报学以及自然语言处理领域常用的 tf-idf 值作为特征。

最终的测试结果见表 4,第一行是本文提出的模型(利用预训练的词向量初始化词嵌入层,并且在分类器训练过程中不进行调整,如表 3 第一行所示),其后是传统文本分类模型的测试结果。从实验结果可以看出,本文所提出的模型相较于传统的文本分类模型有较为明显的优势,各项评测结果均得到了最好的效果。此外,逻辑斯谛回归模型的准确率、召回率以及 F1 值在传统模型中取得了最好的结果,尤其是召回率相较于其他几项评测标准与本文提出的模型差距最小。面对文本这类非结构化的数据,特别是在探索型对话的识别分析任务上,卷积神经网络相较于普通文本分类模型取得了巨大优势,且在保证效果的同时,卷积神经网络模型不需要构造复杂的特征就能够自动对数据中深层次的特征进行学习。

表 4　本文模型与传统文本分类模型的实验结果

模型	准确率	精准率	召回率	F1 值
本文模型	0.94	0.93	0.96	0.94
NB	0.88	0.87	0.88	0.88
LR	0.89	0.86	0.93	0.89
DT	0.87	0.83	0.93	0.88
RF	0.88	0.85	0.91	0.88

5.3 探索型对话自动识别应用分析

从上述实验结果可以看出,在探索型对话自动识别任务中本文提出的模型在F1值以及召回率上取得了较好结果,在MOOC论坛场景下具有较高的实用性。探索型对话自动识别模型作为帮助教师甄别探索型对话和非探索型对话的有力工具,模型的F1值越高对于减轻教师管理论坛压力的帮助就越大。同时考虑到探索型对话在MOOC论坛中的重要作用,探索型对话被识别为非探索型对话相较于非探索型对话误识别为探索型对话对于整个虚拟社区的知识交流与共享具有更大的负面影响,因此较高的召回率在真正的应用中具有重要意义。

此外,MOOC论坛课程覆盖了较多领域,针对不同领域训练分类模型的时候较低的学习成本在大规模应用场景中具有重要意义。本文提出的模型不需要像传统方法一样人工设置复杂的特征,以GloVe方法学习出的单词表征向量为基础,再结合卷积神经网络较强的特征自学习能力,本文模型只需要将原始文本作为输入用较小的学习成本就可以取得较好的自动识别效果。

6 结语

本文利用GolVe模型训练了对语义有较好理解的词向量,并结合能够自动学习文本更深层次特征的卷积神经网络实现了探索型对话的自动识别,改善了传统文本分类方法对语义编码能力不足并且需要构造复杂特征的问题。从实验结果上看,该模型一方面能够准确学习出包含了单词深层次语言学信息的词向量,并且在分类器训练过程中不断调整词向量能够提高模型效果;另一方面,相比于传统的文本分类模型在准确率、精确率、召回率以及F1值上表现出明显优势。随着MOOC在全世界范围内的不断发展,MOOC论坛中具有较高知识密度的探索型对话将会展现出越来越大的价值,自动识别探索型对话对于帮助教师减轻论坛管理负担、激励学生参与知识交流,乃至后续挖掘探索型对话的巨大价值具有重要意义。

本文研究也存在一些局限和不足,对于探索型对话仅仅对其进行了识别,还缺少更进一步的工作,如针对探索型对话中的相关知识点为学生进行推荐,针对话题重复率较高的非探索型对话可以建立相应的自动问答系统等,这些工作可以在今后的研究中考虑。

参考文献:

[1] DANIEL J. Making sense of MOOCs: Musings in a maze of myth, paradox and possibility [EB/OL]. [2017-12-13]. https://jime.open.ac.uk/articles/10.5334/2012-18/.

[2] WEGERIF R, MERCER N. Computers and reasoning through talk in the classroom[J]. Language and education, 1996, 10(1): 47-64.

[3] MERCER N, LITTLETON K. Dialogue and the development of children's thinking: a sociocultural approach[M]. London: Routledge, 2007.

[4] 杨现民, 唐斯斯, 李冀红. 发展教育大数据: 内涵, 价值和挑战[J]. 现代远程教育研究, 2016, 139(1): 50-61.

[5] SIEMENS G, LONG P. Penetrating the fog: analytics in learning and education[J]. Educause review, 2011, 46(5): 31-40.

[6] CLOW D. MOOCs and the funnel of participation[C]//Proceedings of the third international conference on learning analytics and knowledge. Leuven: ACM, 2013: 185-189.

[7] COETZEE D, FOX A, HEARST M A, et al. Should your MOOC forum use a reputation system? [C]//Proceedings of the 17th ACM conference on computer supported cooperative work & social computing, Baltimore, Maryland, USA: ACM, 2014: 1176-1187.

[8] 陈晓美, 贯君, 王福. 虚拟社区信息运动及其规律研究[J]. 图书情报工作, 2016, 60(6): 64-69.

[9] FERGUSON R. The construction of shared knowledge through asynchronous dialogue[D]. Milton Keynes: The Open University, 2009.

[10] 黄维, 赵鹏. 虚拟社区用户知识共享行为影响因素研究[J]. 情报科学, 2016, 34(4): 68-73.

[11] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// The 2014 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[12] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[C]// Proceedings of the IEEE. New York: IEEE, 1998: 2278-2324.

[13] FERGUSON R, WEI Z, HE Y, et al. An evaluation of learning analytics to identify exploratory dialogue in online discussions [C]//Proceedings of the third international conference on learning analytics and knowledge. Leuven: ACM, 2013: 85-93.

[14] SAHAMI M, DUMAIS S, HECKERMAN D, et al. A Bayesian approach to filtering junk e-mail[C]// Learning for text categorization: papers from the 1998 workshop. Palo Alto: Association for the Advancement of Artificial Intelligence, 1998: 98-105.

[15] JOACHIMS T. Transductive inference for text classification using support vector machines[C]// Proceedings of the sixteenth international conference on machine learning. San Francisco: Morgan Kaufmann Publishers, 1999: 200-209.

[16] 夏火松, 甄化春, 张颖烨, 等. 线上商品评论有效性分类专业领域知识模型的构建研究[J]. 情报学报, 2016, 35(9): 946-954.

[17] 余本功, 李婷, 杨颖. 基于多属性加权的社会化问答社区关键词提取方法[J]. 图书情报工作, 2018, 62(5): 132-139.

[18] 王东波, 何琳, 黄水清. 基于支持向量机的先秦诸子典籍自动

分类研究[J]. 图书情报工作, 2017, 61(12): 71 -76.
[19] 夏火松，杨培，熊淦. 基于特征提取改进的在线评论有效性分类模型[J]. 情报学报, 2015, 34(5): 493 -500.
[20] 杜亚楠，刘业政. 基于修正 G2 特征筛选的中文微博情感组合分类[J]. 情报学报, 2016, 35(4): 349 -357.
[21] 祁瑞华，杨德礼，郭旭，等. 基于多层面文体特征的博客作者身份识别研究[J]. 情报学报, 2015, 34(6): 628 -634.
[22] 李晓军，刘怀亮，杜坤. 一种基于复杂网络模型的作者身份识别方法[J]. 图书情报工作, 2015, 59(18): 102 -107.
[23] BRAUD C, DENIS P. Comparing word representations for implicit discourse relation classification[C]//Proceedings of the 2015 Conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2015: 2201 -2211.
[24] TURIAN J, RATINOV L, BENGIO Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th annual meeting of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 384 -394.
[25] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436 -444.
[26] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(8): 2493 -2537.
[27] KIM Y. Convolutional neural networks for sentence classification [C]// Proceedings of EMNLP. Stroudsburg: Association for Computational Linguistics, 2014.
[28] LUND K, BURGESS C. Producing high-dimensional semantic spaces from lexical co-occurrence[J]. Behavior research methods, instruments & computers, 1996, 28(2): 203 -208.
[29] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(2): 1137 -1155.
[30] 安璐，吴林. 融合主题与情感特征的突发事件微博舆情演化分析[J]. 图书情报工作, 2017, 61(15): 120 -129.
[31] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// International conference on learning representations: workshops track. La Jolla: International conference on representation Learning, 2013.
[32] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P, et al. A convolutional neural network for modelling sentences [C]// Meeting of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2014: 655 -665.
[33] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of machine learning research, 2014, 15(1): 1929 -1958.
[34] KINGMA D P, BA J. Adam: a method for stochastic optimization [EB/OL]. [2017 -12 -30]. https://arxiv.org/abs/1412.6980.
[35] 唐晓波，罗颖利. 融入情感差异和用户兴趣的微博转发预测[J]. 图书情报工作, 2017, 61(9): 102 -110.

作者贡献说明：

董庆兴:拟定选题,提出思路,论文撰写与修改;
李华阳:设计模型,实验验证,结果分析,论文撰写与修改;
曹高辉:论文框架设计、对比试验、论文修改;
夏立新:论文构思和修改。

An Exploratory Posts Detecting Method for MOOC Forums Based on Deep Learning

Dong Qingxing[1,2] Li Huayang[3] Cao Gaohui[1] Xia Lixin[1]
[1] School of Information Management, Central China Normal University, Wuhan 430079
[2] Centre for Studies of Information Resources, Wuhan University, Wuhan 430079
[3] Tencent AI Lab, Shenzhen 518057

Abstract: [**Purpose/significance**] Massive Open Online Course (MOOC) forum is an important source to acquire user review data. Automatically detecting exploratory dialogues with high knowledge density from large amounts of unlabeled data and mining its potential value has a significant impact on the improvement of teaching quality and students' mastery of knowledge. [**Method/process**] We proposed a new auto-detecting model based on deep learning, which firstly uses GloVe algorithm to train word embedding to reinforce semantic understanding for texts and then adopts Convolutional Neural Network (CNN) to automatically learn text features and make classifications on exploratory dialogues. An empirical and comparative study was done on the annotated dataset from the online course Introduction to Psychology on the platform of Xuetang. [**Result/conclusion**] Experiment result shows that using the word embedding pretrained by GloVe and fine tune it while training can improve the performance of our model. Our model gets the F1 score of 0.94, which is greatly improved compared with Naive Bayes model (0.88), Logistic Regression model (0.89), Decision Tree model (0.88) and Random Forest model (0.88) and exhibits great practicality with low learning costs.

Keywords: MOOC forum exploratory posts GloVe Convolutional neural network